アノニマイズした行動履歴に基づく行動情報検索システムの提案

川田正明†　　小川克彦††

†慶應義塾大学大学院 政策・メディア研究科

††慶應義塾大学 環境情報学部

E-Mail: kawatan@sfc.keio.ac.jp, ogw@sfc.keio.ac.jp

概要　今日，緯度，経度などのデータが含まれた位置情報付きのコンテンツが増加の傾向にある．このような新しい種類のコンテンツを使った情報検索が行えることで，ユーザにこれまでにない利益がもたらす可能があると考えられる．しかしながら，位置情報付きのコンテンツは個人情報と成る為に扱いには注意しなければならない．そこで，緯度，経度などの実データを，プライバシー保護されたアノニマイズド行動パターンを生成することにより，似た行動をするユーザを探しコンテンツを提供する検索システムを提案する．本システムの評価にあたっては 36 名の被験者の行動履歴を用いて評価した．結果，本システムは場所と時間の実データに対して有効に機能した．

キーワード: SOM，アノニマイズ，行動履歴，プライバシー保護

Anonymized User Behavior-based Information Retrieval System

MASAAKI KAWATA†　KATSUHIKO OGAWA††

†Graduate School of Media and Governance, Keio University

††Faculty of Environmental Information, Keio University

E-Mail: kawatan@sfc.keio.ac.jp, ogw@sfc.keio.ac.jp

Abstract　Recently, we are easy to get location data included the latitude and the longitude using a mobile phone with GPS receiver. And, contents with location data are increasing. We will get a gain if this new type contents can seeks effectively. However, a location data links to problems of privacy.

This research proposes Anonymized User Behavior-based Information Retrieval System. This system provides optimized contents for each user through protected privacy. And, we evaluated this system whether to work effectively with user behavior data. This user behavior data gathered 9,324 records from 36 people. As a result, this system worked effectively with location and time data.

Keywords: Anonymize, Privacy Protection, SOM, User behaviors

1. はじめに

今日，GPS (Global Positioning System)付きの携帯電話に代表されるように，誰でも容易に位置情報を取得可能な環境が整いつつある．例えば，今自分自身が行っていることを一言で投稿するTwitter[1]や，写真投稿の Web サイトである Flickr [2]のようなサービスでは，ユーザ自ら位置情報を積極的に登録する傾向がみられる．

更に，Yahoo!では Fire Eagle [3] (現在プライベートβテスト中)というサービスを開始しようとしている．これは位置情報付きの Twitter と呼ばれているもので最初から位置情報付きで現在行っていることを投稿していくサービスである．

これらは特に，UGC (User-generated Content)と呼ばれる種類のコンテンツに多く見られる傾向があることを付け加えておく．

このような状況の中，ユーザ毎の行動履歴を取得し，行動パターンを分析することで位置情報を扱った検索エンジンを構築することは非常に有効に機能すると考えられる．

しかしながらその半面，緯度・経度などの位置情報は個人情報と成り得るために，公開されることやそのまま送信すると個人が特定される恐れがある．特に，位置情報の連続データは，人の行動履歴となるために，単一の位置情報より個人が特定し易くなるのは明白である．

そこで本研究では，位置情報を含む行動履歴をアノニマイズすることにより安全，かつ行動履歴を用いて行動パターンが近い者を検索し，コンテンツを提供する行動情報検索システムを提案する．

またこの提案するシステムを， 36 人の被験者によって集められた 9324 件の行動履歴のデータと，アノニマイズした行動パターンを比較し評価を行った．

関連研究として，データマイニングを基にした関連情報を検索するための研究も行われているが，ここでは位置情報のプライバシー保護を目的とした関連研究として pawS，Fixer の２つをとり挙げる．

paws[4] は ETH チューリッヒの Marc Langheinrich 氏が提案した，Privacy Policy に基づく位置情報の保護方法である．ユーザが設定した Privacy Policy の Privacy Preference の条件を満たす場合のみ位置情報を公開する．この方法は，ユーザ毎に Privacy Policy の Privacy Preference を設定する必要がある．また，行動情報の検索を行う際，非公開の位置情報があるために検索出来ないデータが存在することになる．

Fixer[5]は，慶應義塾大学の中西健一氏らが提案した位置情報保護のためのフレームワークである．公開する位置情報の粒度を動的に変更することにより，条件下においては非公開が公開に変化する柔軟性を持つ．

Fixter は，pawS のような設定の複雑さは回避している．しかし行動情報を検索する際は，非公開のデータが存在するために検索出来ないデータが存在することになり，有用なデータを取り逃す問題は依然残る．

2. アノニマイズド行動パターン抽出システムの提案

2.1 コンセプト

我々はユーザの行動履歴から安全に行動情報を検索するためのアルゴリズムを述べる．

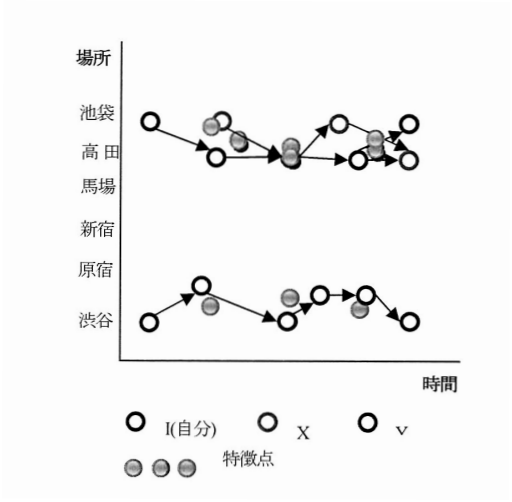

図 1 行動履歴と特徴点の抽出の例

本システムでは，安全ではない緯度・経度などのデータが含まれるユーザの行動履歴から，それぞれのユーザの行動の特徴を代表する「特徴点」を抽出する．このユーザの行動の特徴を捉えた「特徴点」を使い人同士の距離を計量することが可能になる．(図 1 参照)

また，プライバシーの保護の観点からは，ケンブリッジ大学の Alastair Beresford 氏は，「どこでどのくらい時間を過ごしたか」を集計することでユーザを識別することが出来る可能性を述べている．[6]

このような危惧を回避するためにも元のデータからユーザの行動の特徴を表す「特徴点」へ置き換えることで，元のデータを推測されにくくする．

これによりユーザの行動の特徴を捉えつつ安全性を高めるということが出来ると我々は考えている．

尚，この安全性が高められた「特徴点」の集合を「アノニマイズド行動パターン」と定義する．また，本システムは，「アノニマイズド行動パターン抽出システム (Anonymized User Behavior Pattern Extracting System)」と名付けることにする．(図 2 参照)

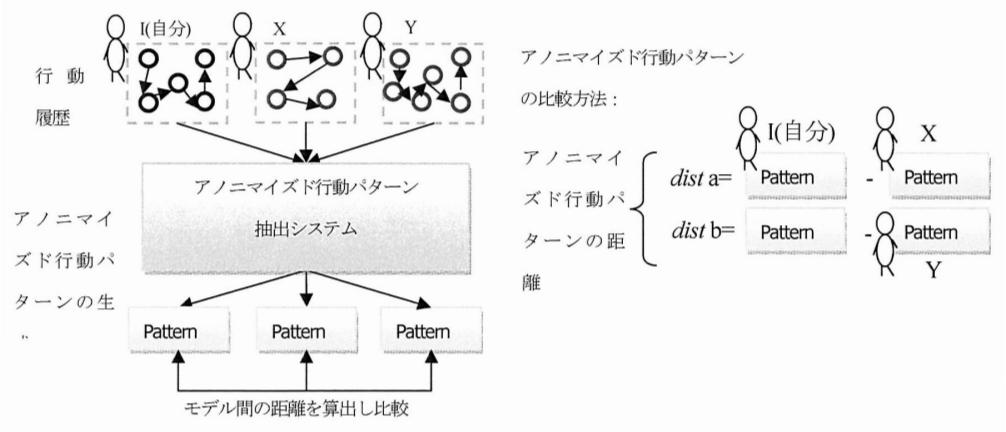

図 2 アノニマイズド行動パターン抽出システム

2.2 SOM によるアノニマイズド行動パターンの抽出

安全性の高い行動の特徴点の集合を抽出したアノニマイズド行動パターンを抽出する方法として，本研究では教師無し学習である SOM (Self-Organizing Maps) [7]を使用した.

この行動モデルを生成するに辺り，我々は緯度(x)・経度(y)・時間(t)・ユーザの性別(g)・年齢(a)という 5 つのデータを SOM に入力する際の入力ベクトルとした．(数式 1 参照)

$$IV_i = \begin{bmatrix} xi \\ yi \\ t_i \\ gi \\ ai \end{bmatrix}$$

数式 1 入力ベクトル

入力ベクトルの要素 xi は，特徴点を鮮明に出すために xi の値が経度の最小値を 0 にし経度の最大値の間に収まるようにした.

long は経度の値, minlong は最小の経度の値, maxlong は最大の経度の値である. (数式 2 参照)

$$xi = (long_i - \min long)/(\max long - \min long)$$

数式 2 入力ベクトルの要素 xi

入力ベクトルの要素 yi も xi と同じコンセプトである．yi の値が緯度の最小値を 0 にし緯度の最大値の間で収まるようにした.

lat は緯度の値，minlat は最小の緯度の値，maxlat は最大の緯度の値である．(数式 3 参照)

$$yi = (lat_i - \min lat)/(\max lat - \min lat)$$

数式 3 入力ベクトルの要素 yi

入力ベクトルの要素 ti は，時と分だけを考慮した.

hours が時，minutes が分の値である．(数式 4 参照)

$$t_i = (hours_i \cdot 60 + minutes_i)/1440$$

数式 4 入力ベクトルの要素 ti

入力ベクトルの要素 gi は，男性が 0，女性が 1 に設定される．(数式 5 参照)

$$\begin{cases} 0, \text{if } gender \text{ is male} \\ 1, \text{otherwise} \end{cases}$$

数式 5 入力ベクトルの要素 gi

入力ベクトル ai は，特徴点を鮮明に出すために最少の年齢を 0 とし，最大の年齢の値の間で値が収まるようにした.

age が年齢，minage が最小の年齢，maxage が最大の年齢である．(数式 6 参照)

$$ai = (age_i - \min age)/(\max age - \min age)$$

数式 6 入力ベクトルの要素 ai

この入力ベクトル群を基に T.Kohonen の学習則を用いて出力ベクトル OV を得る．この OV がアノニマイズド行動パターンとなる．また，OV に含まれる 1 つのベクトルデータが「特徴点」となる．尚，T.Kohonen.の学習則を用いる前に入力ベクトル群は時間 t をキーとしてソートする．つまり，時系列順に揃える.

T.Kohonen の学習則を適用して出力されたアノニマイズド行動パターン OV と別のアノニマイズド行動パターンを，数式 7 を用いて距離を求める．

$$dist = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} \left(OV_i a_j - OV_i b_j\right)^2}$$

数式 7 アノニマイズド行動パターン間の距離

ここで得られる行動パターン同士の距離で，緯度・経度などの実データを用いることなくユーザ間の行動がどれだけ似ているか，または似ていないか計量することが可能になる．

2.3 システムの構築

実際に，アノニマイズド行動パターンを生成し，行動パターン同士を計量し比較する一連のシステムの構築を行った．

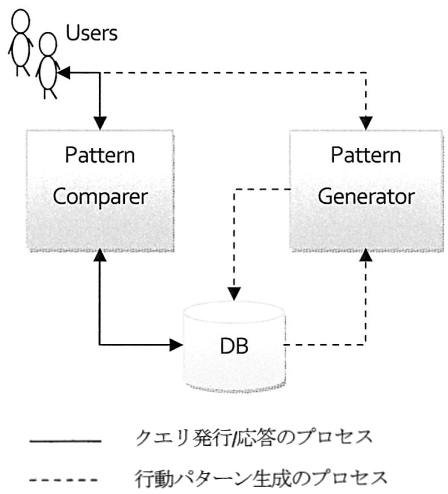

図 3 システムの構成

本システムでは大きく 3 つのビルディングブロックから成る．

それは行動パターンを生成する「Pattern Generator」とユーザからのクエリを受け比較後，結果を返す「Pattern Comparer」，そしてデータの保存と読み込み用に「DB」である．尚，Pattern Generator と Pattern Compare は C#と C++言語で実装した．DB は Microsoft SQL Server を用いた．

そしてこのシステムは大きく分けると 2 つの動作を行う．それは，行動パターンの生成と行動パターンの比較である．(図 3 参照)

行動パターンの生成では，ユーザからの行動履歴，または DB に保存されている行動履歴を使い「Pattern Generator」が行動パターンを生成する．その後，行動履歴を DB から読み込んだ場合には DB に保存される．ユーザから得た行動履歴の場合には，まず行動パターンは DB に保存され，次に行動パターンがユーザへ返される．

次に，行動パターンの比較だが，行動パターンをクエリとしてユーザが発行し「Pattern Comparer」が DB に保存されている行動パターンを比較し行動パターン同士の距離を求める．

この距離をキーにソートしたものを近いものから順にユーザに返す．つまり，クエリとして発行された行動パターンと似ている別の行動パターンがユーザに返される．このような似た行動パターンを探す以外に，似ていない行動パターンを探すことも逆に可能である．

3. 結果

アノニマイズド行動パターン抽出システムを構築し評価した．

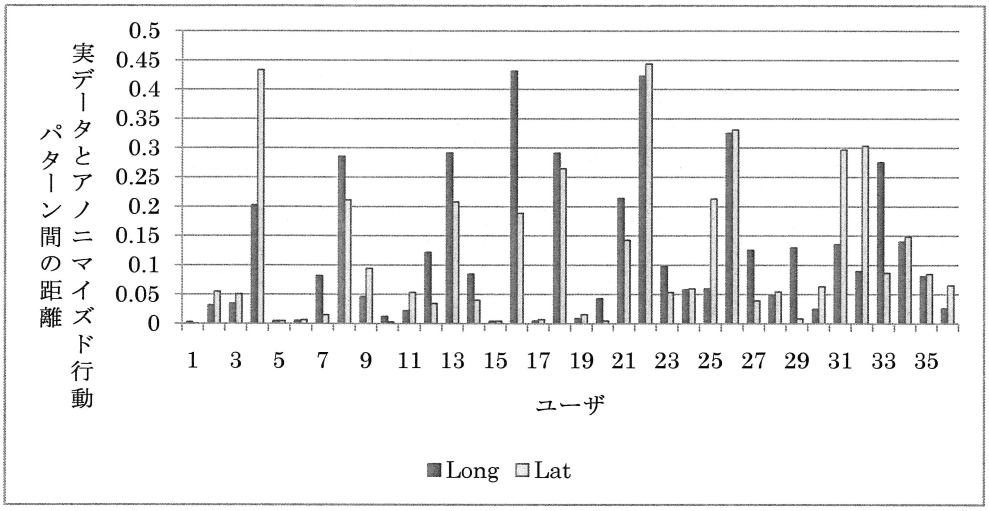

図 4 場所(緯度・経度)の実データとアノニマイズド行動パターンの各要素間の距離

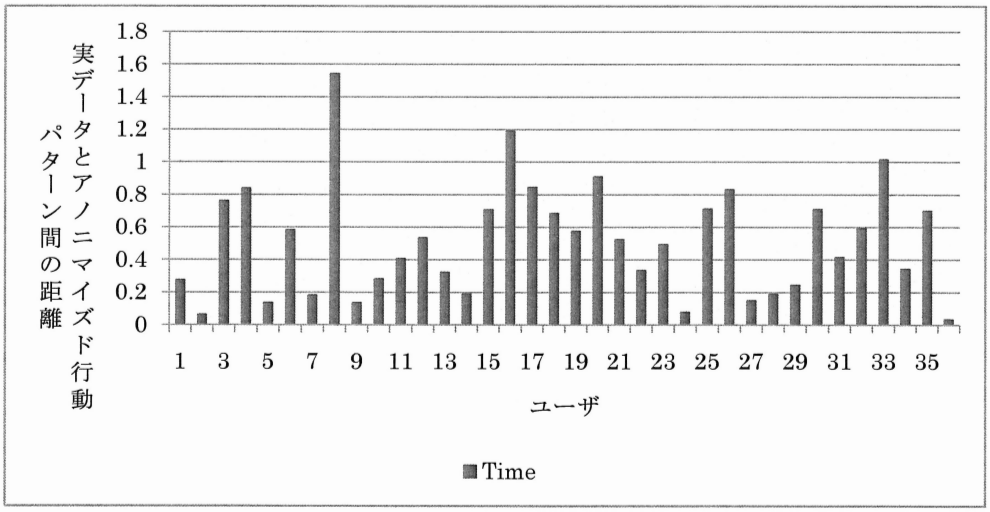

図 5 時間の実データとアノニマイズド行動パターンの各要素間の距離

評価内容は，実データとアノニマイズされた行動パターンの距離の計測した．この計測によりどれだけ実データからアノニマイズされているのかを評価することが可能である．

つまり，距離が離れていればいるほどアノニマイズされていることになる．

3.1 被験者情報と収集したデータ

この評価には情報大航海プロジェクト[8]で

収集した公開データ[9]を用いた．このデータは被験者数 36 人から集めた 9324 件の位置情報付きの行動履歴である．

年齢は 20 歳代から 50 歳代まで，そして 90 歳代 1 人を含む 36 人である．また，全体で男性は 19 人，女性は 17 人という構成である．(表 1 参照)

表 1 被験者の内訳

	20 歳代	30 歳代	40 歳代	50 歳代	90 歳代	合計
男性	6	5	5	2	1	19
女性	7	4	3	3	0	17
合計	13	9	8	5	1	36

3.2 検証：実データとアノニマイズド行動パターン間の距離はどのくらいあるのか

このデータを基に合計 36 個のアノニマイズド行動パターンを生成し実データ間の距離を求めた．この時，アノニマイズド行動パターンに含まれる特徴点の数は全て 10 個に設定した．

この特徴点の数の設定については，1 日の行動履歴が平均して 10 個程度であることから 10 個に決定した．

実データとアノニマイズド行動パターンを，ユークリッド距離を用いて各要素の距離を計測した結果が図 4 と図 5 となっている．

この中で最大値は時間の 1.54243 で，最小値は性別と年齢を省くと緯度の 0.0027 であった．0.00268 という値でも緯度で考えれば差は大きいといえる．性別と年齢は常に変化がないため差は 0 に近いものになった．(表 2 参照)

表 2 実データとアノニマイズド行動パターンとの距離

	Long	Lat	Time	Gender	Age
MIN	0.00268	0.0004	0.03447	0	0
MAX	0.43135	0.44387	1.54243	0	0
AVG	0.11851	0.11355	0.51541	0	0

このことから，緯度，経度，時間に関しては今回の方法によりアノニマイズが有効に機能していることが伺える．しかし，性別と年齢に関してはこの方法であるとアノニマイズには向いていないことが分かる．

4. まとめと今後の課題

今回，実データである緯度・経度，時間，性別，年齢が含まれたユーザの行動履歴から，アノニマイズされたユーザの特徴を示す行動パターンを生成した．この生成した行動パターンからユーザの行動情報を検索するシステムの構築を行い，実データとアノニマイズされた行動パターンがどれだけアノニマイズされているか評価した．

この評価から，今回，我々が提案したアノニマイズド行動パターン抽出システムの有効性がある程度示されたと考えている．

つまり，値が変化し易い緯度・経度，時間のデータについてはアノニマイズ可能なことが分かった．逆に，値が変化しない性別，変化しにくい年齢などについては，アノニマイズは難しいことが解った．

今後の課題として，今回評価していない行動情報の検索について評価する必要があると言え

る．特に検索精度の評価がある．今回は抽出する特徴点の数は固定させていたが変化した際の検索精度について十分な評価が必要である．

また，アノニマイズされた行動パターンを生成するために，今回使用した5つの要素以外，または5つの中限定された要素の中で検索精度が最も高くなる最適な要素についても検討していく必要がある．

更には，今後アノニマイズド行動パターン生成システムが基盤となり，その上でアプリケーションが開発出来るように取り組んでいきたい．

参考文献

[1] Twitter, http://www.twitter.com/

[2] Flickr, http://www.flickr.com/

[3] Fire Eagle, http://fireeagle.yahoo.net/

[4] Langheinrich, M.: A Privacy Awareness System for Ubiquitous Computing Environment, Ubicomp 2002 Proceeding, Lecture Notes in Computer Science, Vol. 2498, Springer-Verlag, pp. 237-245 (2002).

[5] 中西健一，高汐一紀，徳田英幸，粒度の動的変更による位置匿名性についての考察，情報処理学会 論文誌 Vol.46 (9) 2005年9月 pp.2260-2268

[6] Alastair R. Beresford and Frank Stajano. Location privacy in pervasive computing. IEEE Pervasive Computing, 2(1), 2003.

[7] T. Kohonen, Self-Organizing Maps, Berlin, Germany, 1995, Springer-Verlag.

[8] 情報大航海プロジェクト，http://www2.igvpj.jp/

[9] 天笠邦一，加藤文俊，岡部大介，シチュエーション情報の収集とその活用可能性に関する一考察，第18回 情報処理学会ユビキタスコンピューティングシステム研究会 (投稿中)